Debiasing Pre-trained Contextualised Embeddings

Masahiro Kaneko
Tokyo Metropolitan University
kaneko-masahiro@ed.tmu.ac.jp

Danushka Bollegala*
University of Liverpool, Amazon
danushka@liverpool.ac.uk

Abstract

In comparison to the numerous debiasing methods proposed for the static non-contextualised word embeddings, the discriminative biases in contextualised embeddings have received relatively little attention. We propose a fine-tuning method that can be applied at token- or sentence-levels to debias pre-trained contextualised embeddings. Our proposed method can be applied to any pre-trained contextualised embedding model, without requiring to retrain those models. Using gender bias as an illustrative example, we then conduct a systematic study using several state-of-the-art (SoTA) contextualised representations on multiple benchmark datasets to evaluate the level of biases encoded in different contextualised embeddings before and after debiasing using the proposed method. We find that applying token-level debiasing for all tokens and across all layers of a contextualised embedding model produces the best performance. Interestingly, we observe that there is a trade-off between creating an accurate vs. unbiased contextualised embedding model, and different contextualised embedding models respond differently to this trade-off.

1 Introduction

Contextualised word embeddings have significantly improved performance in numerous natural language processing (NLP) applications (Devlin et al., 2019; Liu et al., 2019; Clark et al., 2020) and have established as the de facto standard for input text representations. Compared to static word embeddings (Pennington et al., 2014; Mikolov et al., 2013) that represent a word by a single vector in all contexts it occurs, contextualised embeddings use dynamic context dependent vectors for representing a word in a specific context. Unfortunately however, it has been shown that, similar to their non-contextual counterparts, contextualised text embeddings also encode various types of unfair biases (Zhao et al., 2019; Bordia and Bowman, 2019; May et al., 2019; Tan and Celis, 2019; Bommasani et al., 2020; Kurita et al., 2019). This is a worrying situation because such biases can easily propagate to the downstream NLP applications that use contextualised text embeddings.

Different types of unfair and discriminative biases such as gender, racial and religious biases have been observed in static word embeddings (Bolukbasi et al., 2016; Zhao et al., 2018a; Rudinger et al., 2018; Zhao et al., 2018b; Elazar and Goldberg, 2018; Kaneko and Bollegala, 2019). As discussed later in § 2 different methods have been proposed for debiasing static word embeddings such as projection-based methods (Kaneko and Bollegala, 2019; Zhao et al., 2018b; Bolukbasi et al., 2016; Ravfogel et al., 2020) and adversarial methods (Xie et al., 2017; Gonen and Goldberg, 2019). In contrast, despite multiple studies reporting that contextualised embeddings to be unfairly biased, methods for debiasing contextualised embeddings are relatively under explored (Dev et al., 2020; Nadeem et al., 2020; Nangia et al., 2020). Compared to static word embeddings, debiasing contextualised embeddings is significantly more challenging due to several reasons as we discuss next.

First, compared to static word embedding models where the semantic representation of a word is limited to a single vector, contextualised embedding models have a significantly large number of parameters related in complex ways. For example, BERT-large model (Devlin et al., 2019) contains 24 layers, 16 attention heads and 340M parameters. Therefore, it is not obvious which parameters are responsible for the unfair biases related to a partic-

*Danushka Bollegala holds concurrent appointments as a Professor at University of Liverpool and as an Amazon Scholar. This paper describes work performed at the University of Liverpool and is not associated with Amazon.

ular word. Because of this reason, projection-based methods, popularly used for debiasing pre-trained static word embeddings, cannot be directly applied to debias pre-trained contextualised word embeddings.

Second, in the case of contextualised embeddings, the biases associated with a particular word's representation is a function of both the target word itself and the context in which it occurs. Therefore, the same word can show unfair biases in some contexts and not in the others. It is important to consider the words that co-occur with the target word in different contexts when debiasing a contextualised embedding model.

Third, pre-training large-scale contextualised embeddings from scratch is time consuming and require specialised hardware such as GPU/TPU clusters. On the other hand, fine-tuning a pre-trained contextualised embedding model for a particular task (possibly using labelled data for the target task) is relatively less expensive. Consequently, the standard practice in the NLP community has been to share[1] pre-trained contextualised embedding models and fine-tune as needed. Therefore, it is desirable that a debiasing method proposed for contextualised embedding models can be applied as a fine-tuning method. In this view, counterfactual data augmentation methods (Zmigrod et al., 2019; Hall Maudslay et al., 2019; Zhao et al., 2019) that swap gender pronouns in the training corpus for creating a gender balanced version of the training data are less attractive when debiasing contextualised embeddings because we must retrain those models on the balanced corpora, which is more expensive compared to fine-tuning.

Using gender-bias as a running example, we address the above-mentioned challenges by proposing a debiasing method that fine-tunes pre-trained contextualised word embeddings[2]. Our proposed method retains the semantic information learnt by the contextualised embedding model with respect to gender-related words, while simultaneously removing any stereotypical biases in the pre-trained model. In particular, our proposed method is agnostic to the internal architecture of the contextualised embedding method and we apply it to debias different pre-trained embeddings such as BERT, RoBERTa (Liu et al., 2019), ALBERT (Lan et al., 2020), DistilBERT (Sanh et al., 2019) and ELECTRA (Clark et al., 2020). Moreover, our proposed method can be applied at token-level or at sentence-level, enabling us to debias at different granularities and on different layers in the pre-trained contextualised embedding model.

Following prior work, we compare the proposed debiasing method in two sentence-level tasks: Sentence Encoder Association Test (**SEAT**; May et al., 2019) and Multi-genre co-reference-based Natural Language Inference (**MNLI**; Dev et al., 2020). Experimental results show that the proposed method not only debiases all contextualised word embedding models compared, but also preserves useful semantic information for solving downstream tasks such as sentiment classification (Socher et al., 2013), paraphrase detection (Dolan and Brockett, 2005), semantic textual similarity measurement (Cer et al., 2017), natural language inference (Dagan et al., 2005; Bar-Haim et al., 2006) and solving Winograd schema (Levesque et al., 2012). We consider gender bias as a running example throughout this paper and evaluate the proposed method with respect to its ability to overcome gender bias in contextualised word embeddings, and defer extensions to other types of biases to future work.

2 Related Work

Prior work on debiasing word embeddings can be broadly categorised into two groups depending on whether they consider static or contextualised word embeddings. Although we focus on contextualised embeddings in this paper, we first briefly describe prior work on debiasing static embeddings for completeness of the discussion.

Bias in Static Word Embeddings: Bolukbasi et al. (2016) proposed a post-processing approach that projects gender-neutral words into a subspace, which is orthogonal to the gender direction defined by a list of gender-definitional words. However, their method ignores gender-definitional words during the subsequent debiasing process, and focus only on words that are *not* predicted as gender-definitional by a classifier. Therefore, if the classifier erroneously predicts a stereotypical word as gender-definitional, it would not get debiased. Zhao et al. (2018b) modified the original GloVe (Pennington et al., 2014) objective to learn gender-neutral word embeddings (GN-GloVe) from a given corpus. Unlike the above-

[1] https://huggingface.co/transformers/pretrained_models.html

[2] Code and debiased embeddings: https://github.com/kanekomasahiro/context-debias

mentioned methods, Kaneko and Bollegala (2019) proposed GP-GloVe, a post-processing method to preserve gender-related information with autoencoder (Kaneko and Bollegala, 2020), while removing discriminatory biases from stereotypical cases.

Adversarial learning (Xie et al., 2017; Elazar and Goldberg, 2018; Li et al., 2018) for debiasing first encode the inputs and then two classifiers are jointly trained – one predicting the target task (for which we must ensure high prediction accuracy) and the other for protected attributes (that must not be easily predictable). Elazar and Goldberg (2018) showed that although it is possible to obtain chance-level development-set accuracy for the protected attributes during training, a post-hoc classifier trained on the encoded inputs can still manage to reach substantially high accuracies for the protected attributes. They conclude that adversarial learning alone does not guarantee invariant representations for the protected attributes. Ravfogel et al. (2020) found that iteratively projecting word embeddings to the null space of the gender direction to further improve the debiasing performance.

Benchmarks for biases in Static Embeddings: Word Embedding Association Test (WEAT; Caliskan et al., 2017) quantifies various biases (e.g. gender, race and age) using semantic similarities between word embeddings. Word Association Test (WAT) measures gender bias over a large set of words (Du et al., 2019) by calculating the gender information vector for each word in a word association graph created in the Small World of Words project (SWOWEN; Deyne et al., 2019) by propagating masculine and feminine words via a random walk (Zhou et al., 2003). SemBias dataset (Zhao et al., 2018b) contains three types of word-pairs: (a) **Definition**, a gender-definition word pair (e.g. hero – heroine), (b) **Stereotype**, a gender-stereotype word pair (e.g., manager – secretary) and (c) **None**, two other word-pairs with similar meanings unrelated to gender (e.g., jazz – blues, pencil – pen). It uses the cosine similarity between the gender directional vector, $(\overrightarrow{he} - \overrightarrow{she})$, and the offset vector $(\boldsymbol{a} - \boldsymbol{b})$ for each word pair, (a, b), in each set to measure gender bias. WinoBias (Zhao et al., 2018a) uses the ability to predict gender pronouns with equal probabilities for gender neutral nouns such as occupations as a test for the gender bias in embeddings.

Bias in Contextualised Word Embeddings: May et al. (2019) extended WEAT using templates to create a sentence-level benchmark for evaluating bias called SEAT. In addition to the attributes proposed in WEAT, they proposed two additional bias types: *angry black woman* and *double binds* (when a woman is doing a role that is typically done by a man that woman is seen as arrogant). They show that compared to static embeddings, contextualised embeddings such as BERT, GPT and ELMo are less biased. However, similar to WEAT, SEAT also only has positive predictive ability and cannot detect the absence of a bias. Bommasani et al. (2020) evaluated the bias in contextualised embeddings by first distilling static embeddings from contextualised embeddings and then using WEAT tests for different types of biases such as gender (male, female), racial (White, Hispanic, Asian) and religion (Christianity, Islam). They found that aggregating the contextualised embedding of a particular word in different contexts via averaging to be the best method for creating a static embedding from a contextualised embedding.

Zhao et al. (2019) showed that contextualised ELMo embeddings also learn gender biases present in the training corpus. Moreover, these biases propagate to a downstream coreference resolution task. They showed that data augmentation by swapping gender helps more than neutralisation by a projection. They obtain the embedding of two input sentences with reversed gender from ELMo, and obtain the debiased embedding by averaging them. It can only be applied to feature-based embeddings, so it cannot be applied to fine-tuning based embeddings like BERT. We directly debias the contextual embeddings. Additionally, data augmentation requires re-training of the embeddings, which is often costly compared to fine-tuning. Kurita et al. (2019) created masked templates such as "_ _ is a nurse" and used BERT to predict the masked gender pronouns. They used the log-odds between male and female pronoun predictions as an evaluation measure and showed that BERT to be biased according to it. Karve et al. (2019) learnt conceptor matrices using class definitions in the WEAT and used the negated conceptors to debias ELMo and BERT. Although their method was effective for ELMo, the results on BERT were mixed. This method can only be applied to context-independent vectors, and it requires the creation of static embeddings from BERT and ELMo as a pre-processing step for debi-

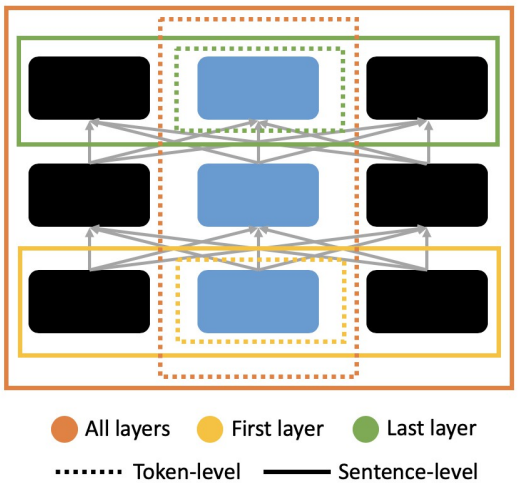


Figure 1: Types of hidden states in E considered in the proposed method. The blue boxes in the middle correspond to the hidden states of the target token.

asing the context-dependent vectors. Therefore, we do not compare against this method in the present study, where we evaluate on context-dependent vectors.

Dev et al. (2020) used natural language inference (NLI) as a bias evaluation task, where the goal is to ascertain if one sentence (i.e. premise) entails or contradictions another (i.e. hypothesis), or if neither conclusions hold (i.e. neutral). The premise-hypothesis pairs are constructed to elicit various types of discriminative biases. They showed that orthogonal projection to gender direction (Dev and Phillips, 2019) can be used to debias contextualised embeddings as well. However, their method can be applied only to the noncontextualised layers (ELMo's Layer 1 and BERT's subtoken layer). In contrast, our proposed method can be applied to *all* layers in a contextualised embedding and outperforms their method on the same NLI task. And our debiasing approach does not require task-dependent data.

3 Debiasing Contextualised Embeddings

We propose a method for debiasing pre-trained contextualised word embeddings in a fine-tuning setting that simultaneously (a) preserves the semantic information in the pre-trained contextualised word embedding model, and (b) removes discriminative gender-related biases via an orthogonal projection in the intermediate (hidden) layers by operating at token- or sentence-levels. Fine-tuning allows debiasing to be carried out without requiring large amounts of tarining data or computational resources. Our debiasing method is independent of model architectures or their pre-training methods, and can be adapted to a wide range of contextualised embeddings as shown in § 4.3.

Let us define two types of words: *attribute* words ($\mathcal{V}_a$) and *target* words ($\mathcal{V}_t$). For example, in the case of gender bias, attribute words consist of multiple word sets such as feminine (e.g. she, woman, her) and masculine (e.g. he, man, him) words, whereas target words can be occupations (e.g. doctor, nurse, professor), which we expect to be gender neutral. We then extract sentences that contain an attribute or a target word. Sentences contain more than one attribute (or target) words are excluded to avoid ambiguities. Let us denote the set of sentences extracted for an attribute or a target word w by $\Omega(w)$. Moreover, let $\mathcal{A} = \bigcup_{w \in \mathcal{V}_a} \Omega(w)$ and $\mathcal{T} = \bigcup_{w \in \mathcal{V}_t} \Omega(w)$ be the sets of sentences containing respectively all of the attribute and target words. We require that the debiased contextualised word embeddings preserve semantic information w.r.t. the sentences in $\mathcal{A}$, and remove any discriminative biases w.r.t. the sentences in $\mathcal{T}$.

Let us consider a contextualised word embedding model E, with pre-trained model parameters $\boldsymbol{\theta}_e$. For an input sentence x, let us denote the embedding of token w in the i-th layer of E by $E_i(w, x; \boldsymbol{\theta}_e)$. Moreover, let the total number of layers in E to be N. In our experiments, we consider different types of encoder models such as E. To formalise the requirement that the debiased word embedding $E_i(t, x; \boldsymbol{\theta}_e)$ of a target word $t \in \mathcal{V}_t$ must not contain any information related to a protected attribute a, we consider the inner-product between the noncontextualised embedding $\boldsymbol{v}_i(a)$ of a and $E_i(t, x; \boldsymbol{\theta}_e)$ as a loss L_i given by (1).

$$L_i = \sum_{t \in \mathcal{V}_t} \sum_{x \in \Omega(t)} \sum_{a \in \mathcal{V}_a} \left(\boldsymbol{v}_i(a)^\top E_i(t, x; \boldsymbol{\theta}_e) \right)^2 \quad (1)$$

Here, $\boldsymbol{v}_i(a)$ is computed by averaging the contextualised embedding of a in the i-th layer of E over all sentences in $\Omega(a)$ following Bommasani et al. (2020) and is given by (2).

$$\boldsymbol{v}_i(a) = \frac{1}{|\Omega(a)|} \sum_{x \in \Omega(a)} E_i(a, x; \boldsymbol{\theta}_e) \quad (2)$$

Here, $|\Omega(a)|$ denotes the total number of sentences in $\Omega(a)$. If a word is split into multiple sub-tokens, we compute the contextualised embedding of the word by averaging the contextualised embeddings

of its constituent sub-tokens. Minimising the loss L_i defined by (1) with respect to $\boldsymbol{\theta}_e$ forces the hidden states of E to be orthogonal to the protected attributes such as gender.

Although removing discriminative biases in E is our main objective, we must ensure that simultaneously we preserve as much useful information that is encoded in the pre-trained model for the downstream tasks. We model this as a regulariser where we measure the squared ℓ_2 distance between the contextualised word embedding of a word w in the i-th layer in the original model, parametrised by $\boldsymbol{\theta}_{\text{pre}}$, and the debiased model as in (3).

$$L_{\text{reg}} = \sum_{x \in \mathcal{A}} \sum_{w \in x} \sum_{i=1}^{N} ||E_i(w, x; \boldsymbol{\theta}_e) - E_i(w, x; \boldsymbol{\theta}_{\text{pre}})||^2 \quad (3)$$

The overall training objective is then given by (4) as the linearly weighted sum of the two losses defined by (1) and (3).

$$L = \alpha L_i + \beta L_{\text{reg}} \quad (4)$$

Here, coefficients $\alpha, \beta \in [0, 1]$ satisfy $\alpha + \beta = 1$.

As shown in Figure 1, a contextualised word embedding model typically contains multiple layers. It is not obvious which hidden states of E are best for calculating L_i for the purpose of debiasing. Therefore, we compute L_i for different layers in a particular contextualised word embedding model in our experiments. Specifically, we consider three settings: debiasing only the **first** layer, **last** layer or **all** layers. Moreover, L_i can be computed only for the target words in a sentence x as in (1), or can be summed up for *all* words in $w \in x$ (i.e. $\sum_{t \in \mathcal{V}_t} \sum_{x \in \Omega(t)} \sum_{w \in x} \left(\boldsymbol{v}_i(a)^\top E_i(w, x; \boldsymbol{\theta}_e)\right)^2$). We refer to the former as **token-level** debiasing and latter **sentence-level** debiasing. Collectively this gives us six different settings for the proposed debiasing method, which we evaluate experimentally in § 4.3.

4 Experiments

4.1 Datasets

We used SEAT (May et al., 2019) 6, 7 and 8 to evaluate gender bias. We use NLI as a downstream evaluation task and use the Multi-Genre Natural Language Inference data (MNLI; Williams et al., 2018) for training and development following Dev et al. (2020). In NLI, the task is to classify a given hypothesis and premise sentence-pair as entailing, contradicting, or neutral. We programmatically generated the evaluation set following Dev et al. (2020) by filling occupation words and gender words in template sentences. The templates take the form "The subject verb a/an object." and the created sentence-pairs are assumed to be neutral.

We used the word lists created by Zhao et al. (2018b) for the attribute list of feminine and masculine words. As for the stereotype word list for target words, we use the list created by Kaneko and Bollegala (2019). Using News-commentary-v15 corpus[3] was extract 11023, 42489 and 34148 sentences respectively for Feminine, Masculine and Stereotype words. We excluded sentences with more than 128 tokens in training data. We randomly sampled 1,000 sentences from each type of extracted sentences as development data.

We used the GLEU benchmark (Wang et al., 2018) to evaluate whether the useful information in the pre-trained embeddings is retrained after debiasing. To evaluate the debiased models with minimal effects due to task-specific fine-tuning, we used the following small-scale training data: Stanford Sentiment Treebank (SST-2; Socher et al., 2013), Microsoft Research Paraphrase Corpus (MRPC; Dolan and Brockett, 2005), Semantic Textual Similarity Benchmark (STS-B; Cer et al., 2017), Recognising Textual Entailment (RTE; Dagan et al., 2005; Bar-Haim et al., 2006; Giampiccolo et al., 2007; Bentivogli et al., 2009), and Winograd Schema Challenge (WNLI; Levesque et al., 2012). We evaluate the performance of the contextualised embeddings on the corresponding development data.

4.2 Hyperparameters

We used BERT (**bert-base-uncased**; Devlin et al., 2019), RoBERTa (**roberta-base**; Liu et al., 2019), ALBERT (**albert-base-v2**; Lan et al., 2020), DistilBERT (**distilbert-base-uncased**; Sanh et al., 2019) and ELECTRA (**electra-small-discriminator**; Clark et al., 2020) in our experiments.[4] DistilBERT has 6 layers and the others 12. We used the development data in SEAT-6 for hyperparameter tuning. The hyperparameters of the models, except the learning rate and batch size, are set to their default values as in `run_glue.py`. Using greedy search, the learning rate was set to 5e-5 and the

[3] http://www.statmt.org/wmt20/translation-task.html

[4] We used https://github.com/huggingface/transformers

Model	*Layer*	*Unit*	SEAT-6	SEAT-7	SEAT-8	#†	SST-2	MRPC	STS-B	RTE	WNLI	Avg
BERT	all	token	$0.68^{\dagger}$	-0.09	$0.60^{\dagger}$	2	92.1	85.6	83.1	60.0	53.5	74.9
		sent	$1.13^{\dagger}$	0.34	0.12	**1**	91.9	82.6	80.0	54.2	40.8	69.9
	last	token	$1.02^{\dagger}$	-1.18	$0.47^{\dagger}$	2	92.2	86.9	82.3	58.1	56.3	75.2
		sent	$1.51^{\dagger}$	-0.60	$1.52^{\dagger}$	2	92.3	84.6	82.9	62.1	56.3	75.6
	first	token	$0.88^{\dagger}$	0.33	$0.86^{\dagger}$	2	92.4	87.1	82.6	62.1	50.7	75.0
		sent	$0.94^{\dagger}$	0.32	$0.97^{\dagger}$	2	91.9	86.1	83.0	63.9	46.5	74.3
	original		$1.04^{\dagger}$	0.18	$0.81^{\dagger}$	2	92.8	86.7	82.4	60.6	56.3	75.8
	random		$1.16^{\dagger}$	-0.08	-0.29	**1**	92.2	87.4	81.9	63.2	54.9	**75.9**
RoBERTa	all	token	$0.51^{\dagger}$	0.15	0.02	**1**	78.1	81.6	73.7	53.8	56.3	68.7
		sent	$1.27^{\dagger}$	$0.86^{\dagger}$	$1.14^{\dagger}$	3	80.3	82.8	74.4	50.9	56.3	68.9
	last	token	$1.17^{\dagger}$	-0.60	$0.45^{\dagger}$	2	79.9	83.7	74.1	52.3	56.3	69.3
		sent	$0.98^{\dagger}$	$0.75^{\dagger}$	$0.87^{\dagger}$	3	69.5	81.5	72.9	52.7	56.3	66.6
	first	token	$1.15^{\dagger}$	0.26	$0.54^{\dagger}$	2	77.8	81.1	74.5	54.5	56.3	68.8
		sent	$1.21^{\dagger}$	0.32	$0.50^{\dagger}$	2	79.0	82.5	74.5	51.6	56.3	68.8
	original		$1.21^{\dagger}$	$1.34^{\dagger}$	$1.01^{\dagger}$	3	93.8	91.2	89.8	71.8	56.3	**80.6**
	random		$1.39^{\dagger}$	$0.40^{\dagger}$	$0.39^{\dagger}$	3	73.4	82.5	73.9	53.4	49.3	66.5
ALBERT	all	token	0.16	0.02	0.18	**0**	78.1	80.5	67.5	54.9	56.3	67,5
		sent	0.18	-0.05	-0.77	**0**	77.3	81.7	69.9	46.9	56.3	66.4
	last	token	$0.83^{\dagger}$	-1.15	-0.76	1	77.8	81.2	68.9	47.3	56.3	66.3
		sent	$0.69^{\dagger}$	-0.06	-0.10	1	78.3	80.1	71.3	55.2	56.3	68,2
	first	token	0.09	0.28	$0.97^{\dagger}$	1	77.9	81.6	70.0	52.0	56.3	67,6
		sent	0.25	$0.60^{\dagger}$	$1.18^{\dagger}$	2	75.9	81.3	70.1	53.1	54.9	67,1
	original		0.30	$0.48^{\dagger}$	$1.12^{\dagger}$	2	92.2	89.9	87.7	70.0	56.3	**79.2**
	random		$0.41^{\dagger}$	0.34	$1.08^{\dagger}$	2	78.2	79.9	71.8	47.3	56.3	66.7
DistilBERT	all	token	$0.70^{\dagger}$	-0.83	-0.66	**1**	90.4	87.8	80.8	56.0	42.3	71.5
		sent	$1.34^{\dagger}$	$1.01^{\dagger}$	$0.97^{\dagger}$	3	91.4	83.3	78.8	57.4	53.5	**72.9**
	last	token	$1.11^{\dagger}$	-0.03	$1.38^{\dagger}$	2	90.9	88.5	80.3	55.6	38.0	70.7
		sent	$1.57^{\dagger}$	-1.34	0.27	**1**	90.8	90.2	80.9	58.5	43.7	72.8
	first	token	$1.19^{\dagger}$	$0.59^{\dagger}$	$0.52^{\dagger}$	3	90.8	90.8	80.4	55.2	38.0	71.0
		sent	$1.19^{\dagger}$	$0.60^{\dagger}$	$0.55^{\dagger}$	3	91.1	90.9	80.1	55.2	36.6	70.8
	original		$1.26^{\dagger}$	0.31	$0.74^{\dagger}$	2	90.8	89.3	80.6	56.0	38.0	70.9
	random		$1.35^{\dagger}$	$0.66^{\dagger}$	-0.25	2	91.1	89.1	80.5	56.3	40.8	71.6
ELECTRA	all	token	0.33	0.10	0.15	**0**	90.3	87.7	79.4	52.7	57.7	**73.6**
		sent	$0.42^{\dagger}$	0.21	0.33	1	90.7	87.1	79.5	52.3	54.9	72.9
	last	token	$0.55^{\dagger}$	0.07	0.24	1	90.8	87.3	79.8	51.6	46.5	71.2
		sent	$0.50^{\dagger}$	$0.42^{\dagger}$	$0.32^{\dagger}$	3	90.5	87.3	80.1	54.5	40.8	70.6
	first	token	0.31	0.10	0.33	**0**	90.4	86.9	79.7	53.1	56.3	73.4
		sent	0.29	0.22	0.30	**0**	90.4	87.6	79.7	53.4	56.3	73.5
	original		0.16	$0.46^{\dagger}$	0.04	1	90.5	87.9	80.4	54.5	46.5	72.0
	random		$0.43^{\dagger}$	$0.49^{\dagger}$	-0.22	2	90.4	87.7	78.5	51.3	54.9	72.6

Table 1: Gender bias of contextualised embeddings on SEAT. † denotes significant bias effects at $\alpha < 0.01$.

batch size to 32 during debiasing. Optimal values for $\alpha = 0.2$ and $\beta = 0.8$ were found by a greedy search in $[0, 1]$ with 0.1 increments. For the GLEU and MNLI experiments, we set the learning rate to 2e-5 and the batch size to 16. Experiments were conducted on a GeForce GTX 1080 Ti GPU.

4.3 Debiasing vs. Preserving Information

Table 1 shows the results on SEAT and GLEU where **original** denotes the pre-trained contextualised models prior to debiasing. We see that original models other than ELECTRA contain significant levels of gender biases. Overall, the **all-token** method that conducts token-level debiasing across all layers performs the best. Prior work has shown that biases are learned at each layer (Bommasani et al., 2020) and it is important to debias all layers. Moreover, we see that debiasing at token-level is more efficient compared to at the sentence-level. This is because in token-level debiasing, the loss is computed only on the target word and provides a more direct debiasing update for the target word than in the sentence-level debiasing, which sums the losses over all tokens in a sentence.

To test the importance of carefully selecting the target words considering the types of biases that we want to remove from the embeddings, we implement a **random** baseline where we randomly select target and attribute words from $\mathcal{V}_a \cup \mathcal{V}_t$ and perform **all-token** debiasing. We see that **random** debiases BERT to some extent but is not effective on other models. This result shows that the

Model	MNLI-m	MNLI-mm	NN	FN	T:0.7
Dev et al. (2020)	**80.8**	81.1	85.5	**97.3**	88.3
all-token	80.7	**81.2**	**87.8**	96.8	**89.3**
original	**80.8**	81.0	82.3	96.4	83.2
random	80.5	81.1	85.8	96.4	87.0

Table 2: Debias results for BERT in MNLI.

proposed debiasing method is *not* merely a regularisation technique that imposes constraints on any arbitrary set of words, but it is essential to carefully select the target words used for debiasing.

The results on GLEU show that BERT, DistilBERT and ELECTRA compared to the **original** embeddings, the debiased embeddings report comparable performances in most settings. This confirms that the proposed debiasing method preserves sufficient semantic information contained in the **original** embeddings that can be used to learn accurate prediction models for the downstream NLP tasks.[5] However, the performance of RoBERTa and ALBERT decrease significantly compared to their **original** versions after debiasing. We suspect that these models are more sensitive to fine-tuning and hence lose their pre-trained information during the debiasing process. We defer the development of techniques to address this issue to future research.

4.4 Measuring Bias with Inference

Following Dev et al. (2020), we use the multi-genre co-reference-based natural language inference (MNLI) dataset for evaluating gender bias. This dataset contains sentence triples where a premise must be neutral in entailment w.r.t. two hypotheses. If the predictions made by a classifier that uses word embeddings as features deviate from neutrality, it is considered as biased. Given a set containing M test instances, let the entailment predictor's probabilities for the m-th instance for entail, neutral and contradiction labels be respectively e_m, n_m and c_m. Then, they proposed the following measures to quantify the bias: (1) Net Neutral (**NN**): $\text{NN} = \frac{1}{M}\sum_{m=1}^{M} n_m$; (2) Fraction Neutral (**FN**): $\text{FN} = \frac{1}{M}\sum_{m=1}^{M} \mathbf{1}[\text{neutral} = \max(e_m, n_m, c_m)]$; and (3) Threshold τ (**T:**τ): T:τ $= \mathbf{1}[n_m \geq \tau]$, where we used $\tau = 0.7$ following Dev et al. (2020). For an ideal (bias-free) embedding, all three measure would be 1.

[5]Although on WNLI **all-token** debiasing improves performance for DistilBERT and ELECTRA compared to the respective **original** models, this is insignificant as WNLI contains only 146 test instances.

Model	*Layer*	SEAT-6	SEAT-7	SEAT-8
BERT	all	**0.44**	0.25	**0.46**
	last	0.56	**0.12**	0.47
	first	0.52	0.22	0.49
RoBERTa	all	**0.59**	**0.23**	0.61
	last	0.73	0.24	0.65
	first	0.69	0.28	**0.59**
ALBERT	all	**0.46**	0.48	**0.24**
	last	1.15	**0.26**	0.60
	first	0.54	0.89	0.95
DistilBERT	all	**0.66**	**-0.16**	0.37
	last	0.88	0.19	**0.35**
	first	0.90	0.40	0.52
ELECTRA	all	**0.21**	**0.02**	**0.18**
	last	0.34	0.20	0.21
	first	0.28	0.13	0.34

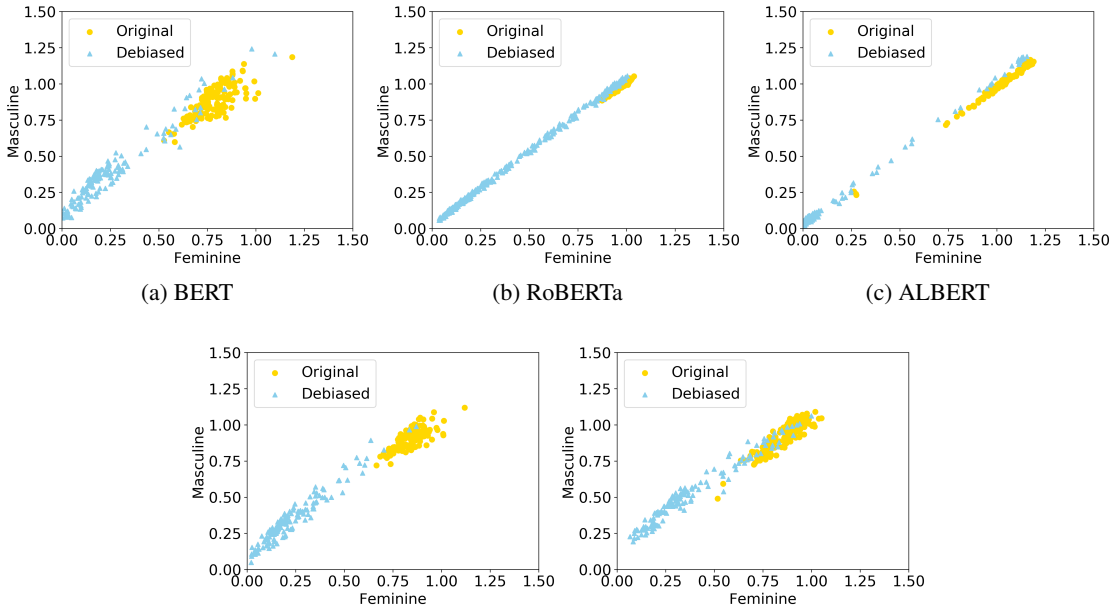
Table 3: Averaged scores over all layers in an embedding debiased at token-level, measured on SEAT tests.

In Table 2, we compare our proposed method against the *noncontextualised debiasing* method proposed by Dev et al. (2020) where they debias Layer 1 of BERT-large model using an orthogonal projection to the gender direction during training and evaluation. In addition to the above-mentioned measures, we also report the entailment accuracy on the matched (in-domain) and mismatched (cross-domain) denoted respectively by **MNLI-m** and **MNLI-mm** in Table 2 to evaluate the semantic information preserved in the embeddings after debiasing.

We see that the proposed method outperforms noncontextualised debiasing (Dev et al., 2020) in NN and T:0.7, and its performance of the MNLI task is comparable to the original embeddings. This result further confirms that the proposed method can not only debias well but can also preserve the pre-trained information. Moreover, it is consistent with the results reported in Table 1 and shows that debiasing all layers is more effective than only the first layer as done by Dev et al. (2020).

4.5 The Importance of Debiasing All Layers

In Table 1, we investigated the bias for the final layer, but it is known that the contextualised embeddings are learned at each layer (Bommasani et al., 2020). Therefore, to investigate whether by debiasing in each layer we are able to remove the biases of the entire contextualised embeddings, we evaluate the debiased embeddings at each layer



(a) BERT (b) RoBERTa (c) ALBERT

(d) DistilBERT (e) ELECTRA

Figure 2: Scatter plot of gender information of hidden states for original and debiased stereotype words.

on SEAT 6, 7, 8 datasets and report the averaged metrics for **all-token**, **first-token** and **last-token** methods in Table 3. We see that, on average, **fitst-token** and **last-token** methods have more bias than **all-token**. Therefore, we conclude that It is not enough to debias only the first and last layers even in DistilBERT, which has a small number of layers. These results show that biases in the entire contextualised embedding cannot be reliably removed by debiasing only some selected layers, but rather the importance of debiasing all layers consistently.

4.6 Visualizing Debiasing Results

To further illustrate the effect of debiasing using the proposed **all-token** method, we visualise the similarity scores of a stereotypical word with feminine and masculine dimensions as follows. First, for each target word t, its hidden state, $E_i(t, x)$ in the i-th layer of the model E in a sentence x is computed. Next, we average those hidden states across all sentences in the dataset that contain t to obtain $\hat{E}_i(t) = \frac{1}{|\mathcal{T}|}\sum_{x \in \mathcal{T}} E_i(t, x)$. Likewise, we compute $\hat{E}_i(f)$ and $\hat{E}_i(m)$ respectively for each feminine (f) and masculine (m) word. Next, we compute, s_i^f, the cosine similarity between each $\hat{E}_i(f)$ and the feminine vector $\boldsymbol{v}_i(f)$, and the cosine similarity, s_i^m, between each $\hat{E}_i(f)$ and the masculine vector $\boldsymbol{v}_i(f)$. $\boldsymbol{s}_i^f$ and $\boldsymbol{s}_i^m$, respectively, are averaged over all layers in a contextualised embedding model to obtain $\boldsymbol{s}^f_{\text{Avg}}$ and $\boldsymbol{s}^m_{\text{Avg}}$, which represent how much gender information each gender word contains on average.

We then compute the cosine similarity, $\boldsymbol{s}_i^{t,f}$, between each stereotype word's averaged embedding, $\hat{E}_i(t)$ and the feminine vector $\boldsymbol{v}_i(f)$. Similarly, we compute the cosine similarity $\boldsymbol{s}_i^{t,m}$ between each stereotype word's averaged embedding $\hat{E}_i(t)$ and the masculine vector $\boldsymbol{v}_i(m)$. We then average $\boldsymbol{s}^{t,f}$ and $\boldsymbol{s}^{t,m}$ over the layers in E respectively, to compute $\boldsymbol{s}^{t,f}_{\text{Avg}}$ and $\boldsymbol{s}^{t,m}_{\text{Avg}}$, which represent how much gender information each stereotype word contains on average. Finally, we visualise the normalised female and male gender scores given respectively by $\boldsymbol{s}^{t,f}_{\text{Avg}}/\boldsymbol{s}^{f}_{\text{Avg}}$ and $\boldsymbol{s}^{t,m}_{\text{Avg}}/\boldsymbol{s}^{m}_{\text{Avg}}$. For example, a zero $\boldsymbol{s}^{t,f}_{\text{Avg}}/\boldsymbol{s}^{f}_{\text{Avg}}$ value indicates that t does not contain female gender related information, whereas a value of one indicates that it contains all information about the female gender. Figure 2 shows each stereotype word with its normalised female ad male gender scores respectively in x and y axises. For a word, a yellow circle denotes its original embeddings, and the blue triangle denotes the result of debiasing using the **all-token** method.

We see that with the original embeddings, stereotypical words of are distributed close to one, indicating that they are highly gender-specific. On the other hand, we see that the debiased BERT, DistilBERT and ELECTRA have similar word distributions compared to the original embeddings respectively, with an overall movement towards

zero. On the other hand, for RoBERTa, debiased embeddings are mainly distributed from zero to around one compared to the original embeddings. Moreover, for ALBERT, the debiased embeddings are close to zero, but unlike the original distribution, the debiased embeddings are mainly clustered around zero. This shows that RoBERTa and ALBERT do not retain structure of the original distribution after debiasing. While ALBERT over-debiases pre-trained embeddings of stereotypical words, RoBERTa under-debiases them. This trend was already confirmed on the downstream evaluation tasks conducted in Table 1.

5 Conclusion

We proposed a debiasing method for pre-trained contextualised word embeddings, operating at token- or sentence-levels. Our experimental results showed that the proposed method effectively debiases discriminative gender-related biases, while preserving useful semantic information in the pre-trained embeddings. The results showed that the downstream task was more effective in debias than the previous studies.

References

Roy Bar-Haim, Ido Dagan, Bill Dolan, Lisa Ferro, and Danilo Giampiccolo. 2006. The second pascal recognising textual entailment challenge. *Proceedings of the Second PASCAL Challenges Workshop on Recognising Textual Entailment*.

Luisa Bentivogli, Ido Dagan, Hoa Trang Dang, Danilo Giampiccolo, and Bernardo Magnini. 2009. The fifth pascal recognizing textual entailment challenge. In *In Proc Text Analysis Conference (TAC'09*.

Tolga Bolukbasi, Kai-Wei Chang, James Y. Zou, Venkatesh Saligrama, and Adam Kalai. 2016. Man is to computer programmer as woman is to homemaker? debiasing word embeddings. In *NIPS*.

Rishi Bommasani, Kelly Davis, and Claire Cardie. 2020. Interpreting Pretrained Contextualized Representations via Reductions to Static Embeddings. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 4758–4781, Online. Association for Computational Linguistics.

Shikha Bordia and Samuel R. Bowman. 2019. Identifying and reducing gender bias in word-level language models. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Student Research Workshop*, pages 7–15, Minneapolis, Minnesota. Association for Computational Linguistics.

Aylin Caliskan, Joanna J Bryson, and Arvind Narayanan. 2017. Semantics derived automatically from language corpora contain human-like biases. *Science*, 356:183–186.

Daniel Cer, Mona Diab, Eneko Agirre, Iñigo Lopez-Gazpio, and Lucia Specia. 2017. SemEval-2017 task 1: Semantic textual similarity multilingual and crosslingual focused evaluation. In *Proceedings of the 11th International Workshop on Semantic Evaluation (SemEval-2017)*, pages 1–14, Vancouver, Canada. Association for Computational Linguistics.

K. Clark, Minh-Thang Luong, Quoc V. Le, and Christopher D. Manning. 2020. Electra: Pre-training text encoders as discriminators rather than generators. *ArXiv*, abs/2003.10555.

Ido Dagan, Oren Glickman, and Bernardo Magnini. 2005. The pascal recognising textual entailment challenge. In *Proceedings of the First International Conference on Machine Learning Challenges: Evaluating Predictive Uncertainty Visual Object Classification, and Recognizing Textual Entailment*, MLCW'05, pages 177–190, Berlin, Heidelberg. Springer-Verlag.

Sunipa Dev, Tao Li, Jeff Phillips, and Vivek Srikumar. 2020. On Measuring and Mitigating Biased Inferences of Word Embeddings. In *AAAI*.

Sunipa Dev and Jeff M. Phillips. 2019. Attenuating bias in word vectors. In *The 22nd International Conference on Artificial Intelligence and Statistics, AISTATS 2019, 16-18 April 2019, Naha, Okinawa, Japan*, volume 89 of *Proceedings of Machine Learning Research*, pages 879–887. PMLR.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. BERT: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 4171–4186, Minneapolis, Minnesota. Association for Computational Linguistics.

Simon De Deyne, Danielle J. Navarro, Amy Perfors, Marc Brysbaert, and Gert Storms. 2019. The "small world of words" english word association norms for over 12,000 cue words. *Behavior Research Methods*, 51:987–1006.

William B. Dolan and Chris Brockett. 2005. Automatically constructing a corpus of sentential paraphrases. In *Proceedings of the Third International Workshop on Paraphrasing (IWP2005)*.

Yupei Du, Yuanbin Wu, and Man Lan. 2019. Exploring human gender stereotypes with word association test. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 6133–6143, Hong Kong, China. Association for Computational Linguistics.

Yanai Elazar and Yoav Goldberg. 2018. Adversarial Removal of Demographic Attributes from Text Data. In *Proc. of EMNLP*.

Danilo Giampiccolo, Bernardo Magnini, Ido Dagan, and Bill Dolan. 2007. *The Third PASCAL Recognizing Textual Entailment Challenge*, pages 1–9. Association for Computational Linguistics, USA.

Hila Gonen and Yoav Goldberg. 2019. Lipstick on a pig: Debiasing methods cover up systematic gender biases in word embeddings but do not remove them. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 609–614, Minneapolis, Minnesota. Association for Computational Linguistics.

Rowan Hall Maudslay, Hila Gonen, Ryan Cotterell, and Simone Teufel. 2019. It's all in the name: Mitigating gender bias with name-based counterfactual data substitution. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 5266–5274, Hong Kong, China. Association for Computational Linguistics.

Masahiro Kaneko and Danushka Bollegala. 2019. Gender-preserving debiasing for pre-trained word embeddings. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 1641–1650, Florence, Italy. Association for Computational Linguistics.

Masahiro Kaneko and Danushka Bollegala. 2020. Autoencoding improves pre-trained word embeddings. In *Proceedings of the 28th International Conference on Computational Linguistics*, pages 1699–1713, Barcelona, Spain (Online). International Committee on Computational Linguistics.

Saket Karve, Lyle Ungar, and João Sedoc. 2019. Conceptor debiasing of word representations evaluated on WEAT. In *Proceedings of the First Workshop on Gender Bias in Natural Language Processing*, pages 40–48, Florence, Italy. Association for Computational Linguistics.

Keita Kurita, Nidhi Vyas, Ayush Pareek, Alan W Black, and Yulia Tsvetkov. 2019. Measuring bias in contextualized word representations. In *Proceedings of the First Workshop on Gender Bias in Natural Language Processing*, pages 166–172, Florence, Italy. Association for Computational Linguistics.

Zhenzhong Lan, Mingda Chen, Sebastian Goodman, Kevin Gimpel, Piyush Sharma, and Radu Soricut. 2020. Albert: A lite bert for self-supervised learning of language representations. *ArXiv*, abs/1909.11942.

Hector J. Levesque, Ernest Davis, and Leora Morgenstern. 2012. The winograd schema challenge. In *Proceedings of the Thirteenth International Conference on Principles of Knowledge Representation and Reasoning*, KR'12, pages 552–561. AAAI Press.

Yitong Li, Timothy Baldwin, and Trevor Cohn. 2018. Towards robust and privacy-preserving text representations. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 25–30. Association for Computational Linguistics.

Yinhan Liu, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, Mike Lewis, Luke Zettlemoyer, and Veselin Stoyanov. 2019. Roberta: A robustly optimized bert pretraining approach. *arXiv preprint arXiv:1907.11692*.

Chandler May, Alex Wang, Shikha Bordia, Samuel R. Bowman, and Rachel Rudinger. 2019. On measuring social biases in sentence encoders. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 622–628, Minneapolis, Minnesota. Association for Computational Linguistics.

Tomas Mikolov, Kai Chen, and Jeffrey Dean. 2013. Efficient estimation of word representation in vector space. In *ICLR*.

Moin Nadeem, Anna Bethke, and Siva Reddy. 2020. StereoSet: Measuring stereotypical bias in pretrained language models.

Nikita Nangia, Clara Vania, Rasika Bhalerao, and Samuel R. Bowman. 2020. CrowS-Pairs: A Challenge Dataset for Measuring Social Biases in Masked Language Models. In *Proc. of EMNLP*.

Jeffery Pennington, Richard Socher, and Christopher D. Manning. 2014. Glove: global vectors for word representation. In *EMNLP*, pages 1532–1543.

Shauli Ravfogel, Yanai Elazar, Hila Gonen, Michael Twiton, and Yoav Goldberg. 2020. Null It Out: Guarding Protected Attributes by Iterative Nullspace Projection. In *Proc. of ACL*.

Rachel Rudinger, Jason Naradowsky, Brian Leonard, and Benjamin Van Durme. 2018. Gender bias in coreference resolution. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 2 (Short Papers)*, pages 8–14. Association for Computational Linguistics.

Victor Sanh, Lysandre Debut, Julien Chaumond, and Thomas Wolf. 2019. Distilbert, a distilled version of bert: smaller, faster, cheaper and lighter. *ArXiv*, abs/1910.01108.

Richard Socher, Alex Perelygin, Jean Wu, Jason Chuang, Christopher D. Manning, Andrew Ng, and Christopher Potts. 2013. Recursive deep models for semantic compositionality over a sentiment treebank. In *Proceedings of the 2013 Conference on Empirical Methods in Natural Language Processing*, pages 1631–1642, Seattle, Washington, USA. Association for Computational Linguistics.

Yi Chern Tan and L. Elisa Celis. 2019. Assessing social and intersectional biases in contextualized word representations. In *Advances in Neural Information Processing Systems 32*, pages 13230–13241. Curran Associates, Inc.

Alex Wang, Amanpreet Singh, Julian Michael, Felix Hill, Omer Levy, and Samuel Bowman. 2018. GLUE: A multi-task benchmark and analysis platform for natural language understanding. In *Proceedings of the 2018 EMNLP Workshop BlackboxNLP: Analyzing and Interpreting Neural Networks for NLP*, pages 353–355, Brussels, Belgium. Association for Computational Linguistics.

Adina Williams, Nikita Nangia, and Samuel Bowman. 2018. A broad-coverage challenge corpus for sentence understanding through inference. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers)*, pages 1112–1122, New Orleans, Louisiana. Association for Computational Linguistics.

Qizhe Xie, Zihang Dai, Yulun Du, Eduard Hovy, and Graham Neubig. 2017. Controllable invariance through adversarial feature learning. In *Proc. of NIPS*.

Jieyu Zhao, Tianlu Wang, Mark Yatskar, Ryan Cotterell, Vicente Ordonez, and Kai-Wei Chang. 2019. Gender bias in contextualized word embeddings. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 629–634, Minneapolis, Minnesota. Association for Computational Linguistics.

Jieyu Zhao, Tianlu Wang, Mark Yatskar, Vicente Ordonez, and Kai-Wei Chang. 2018a. Gender bias in coreference resolution: Evaluation and debiasing methods. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 2 (Short Papers)*, pages 15–20. Association for Computational Linguistics.

Jieyu Zhao, Yichao Zhou, Zeyu Li, Wei Wang, and Kai-Wei Chang. 2018b. Learning Gender-Neutral Word Embeddings. In *Proc. of EMNLP*, pages 4847–4853.

Dengyong Zhou, Olivier Bousquet, Thomas Navin Lal, Jason Weston, and Bernhard Schölkopf. 2003. Learning with local and global consistency. In *NIPS*.

Ran Zmigrod, Sebastian J. Mielke, Hanna Wallach, and Ryan Cotterell. 2019. Counterfactual Data Augmentation for Mitigating Gender Stereotypes in Languages with Rich Morphology. In *Proc. of ACL*.